Speed Dating

...

Alejandro Salvat
Maxime
Daniel
Pol Pérez

Index

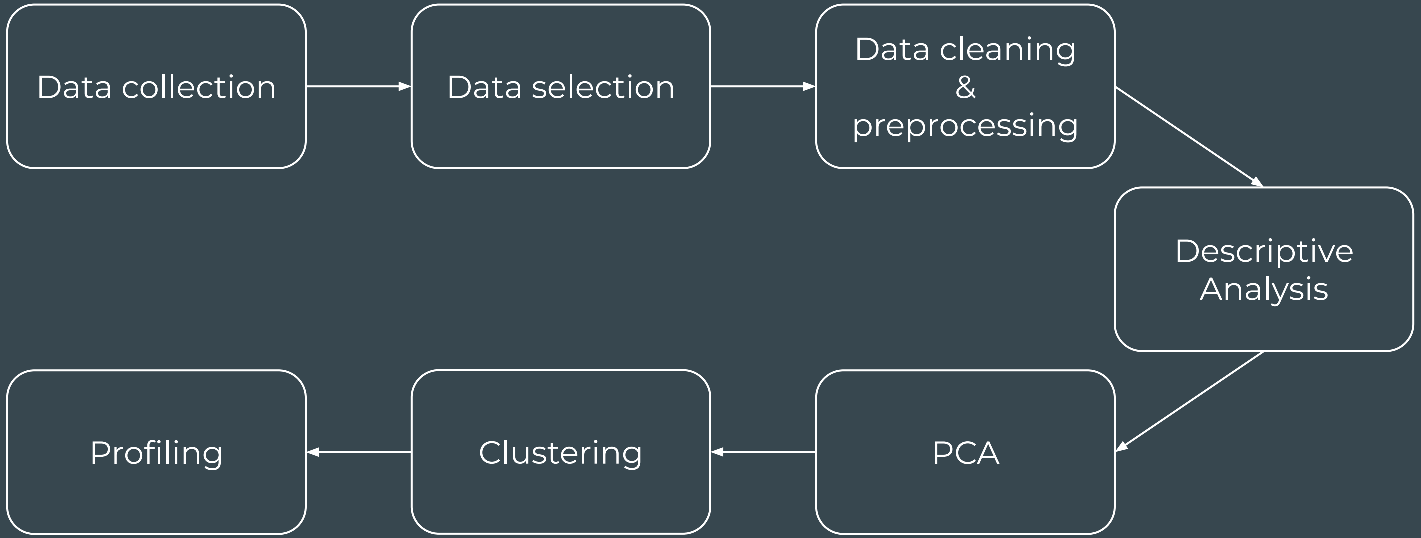
1. Data Preprocessing
    a. Data cleaning
    b. Imputation of null values
    c. Errors values
    d. Outliers treatment
2. Descriptive analysis
3. PCA
4. Clustering
5. Profiling
6. Conclusions

Topics, goals and urls

- Is there any separability in how much you have been liked among the participants?
- Which attributes of a partner do participants attach most importance to?

https://www.kaggle.com/datasets/ulrikthygepedersen/speed-dating

Data mining process



Data cleaning

Reducing the size of our dataset

8378 x 123

↓

8378 x 25

Changing the format of the values

d’0’ → 0
d’1’ → 1

[0...1] → 1
[1...2] → 2
[2...3] → 3
.
.
.
[9...10] → 10

Erase variables with unacceptable number of null values

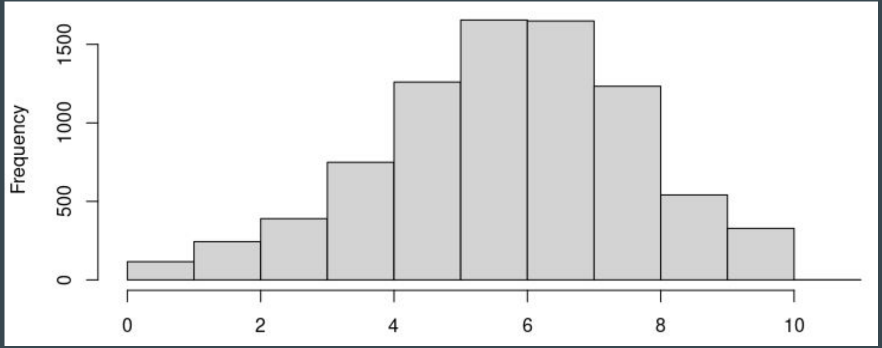
expected_num_interested_in_me

79%
NULL

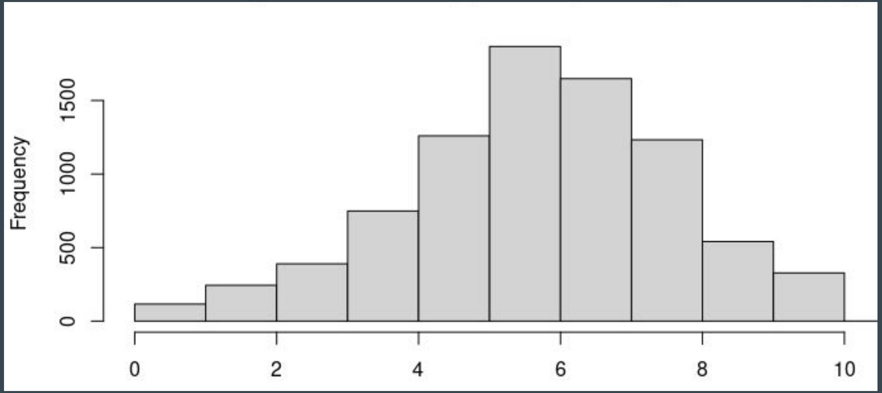
Preprocessing - Imputation of null values

Numerical variables: Mean Imputation

Histogram without imputation



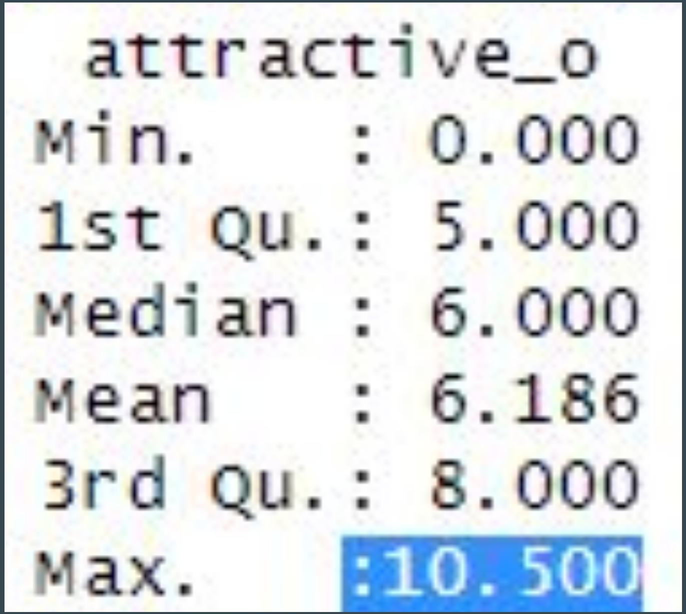
Histogram with mean imputation



Categorical variables: MCA Imputation

Preprocessing - Errors in values

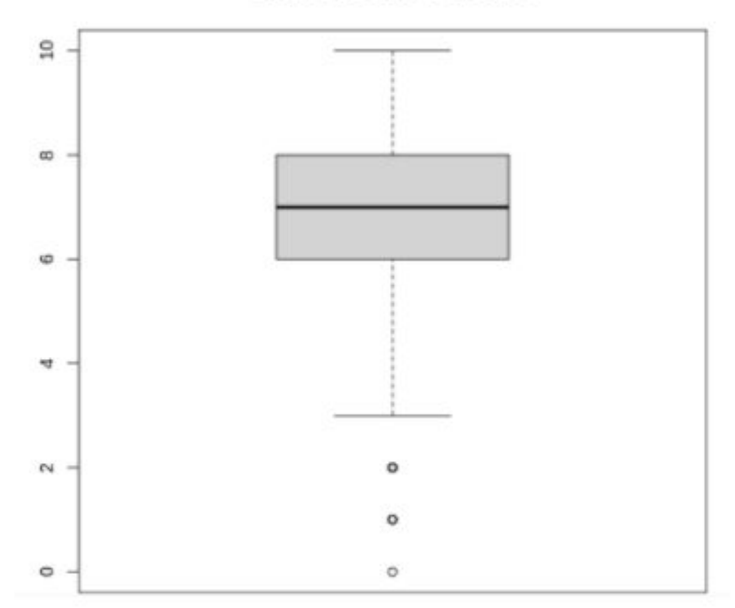
- Are the variables in the range that they should be according to the metadatafile?



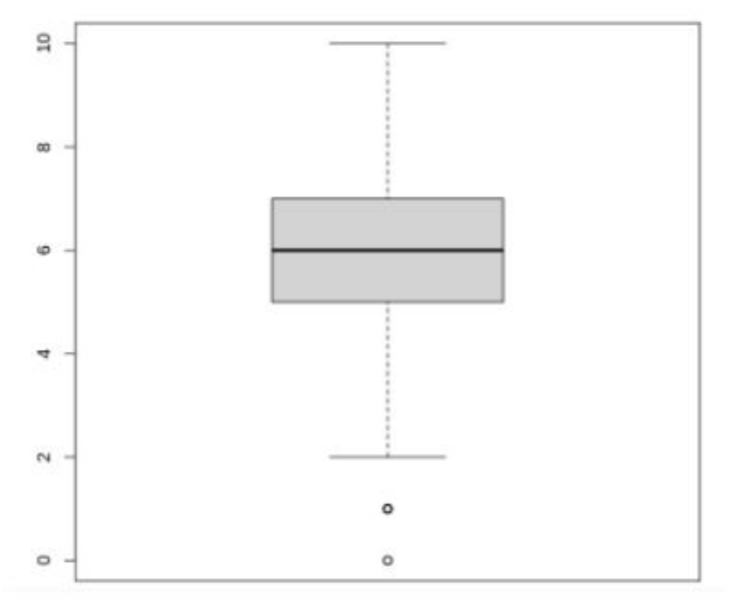
```
 attractive_o
Min.    : 0.000
1st Qu.: 5.000
Median : 6.000
Mean   : 6.186
3rd Qu.: 8.000
Max.   :10.500
```

Preprocessing - Outliers treatment

BoxPlot “sincere_partner”

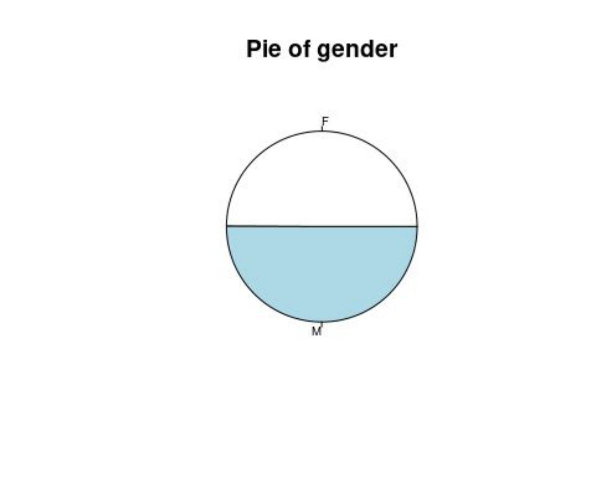


BoxPlot “like”



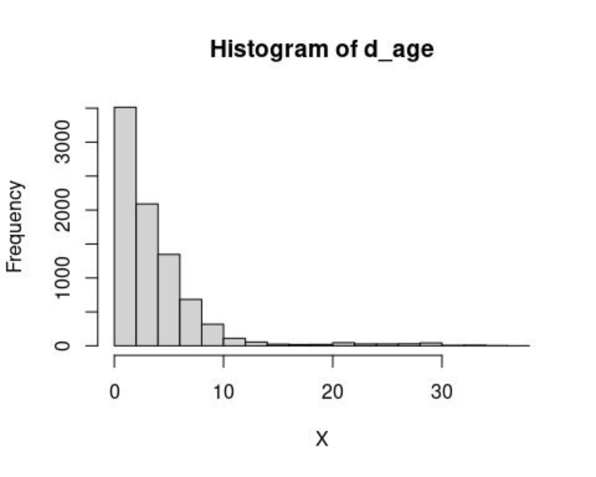
Descriptive analysis

Gender at birth



Binary

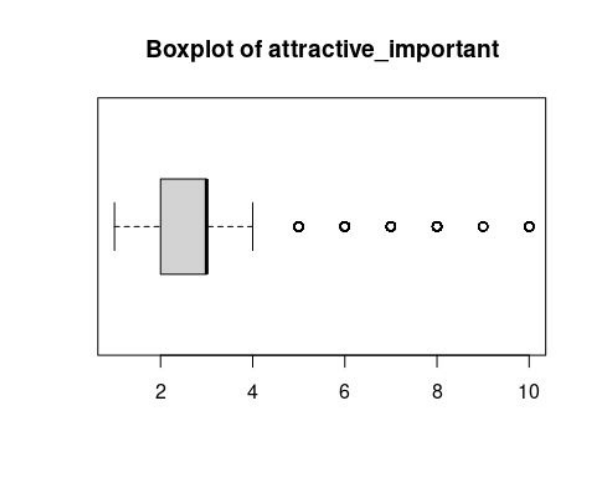
Age difference



Numerical

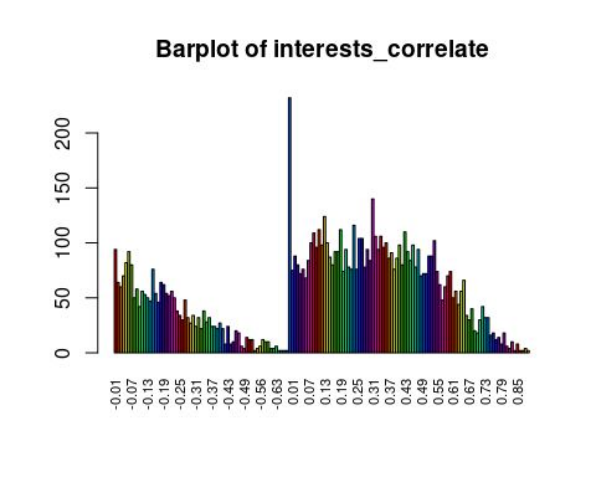
Descriptive analysis

Importance of attractiveness



Categorical

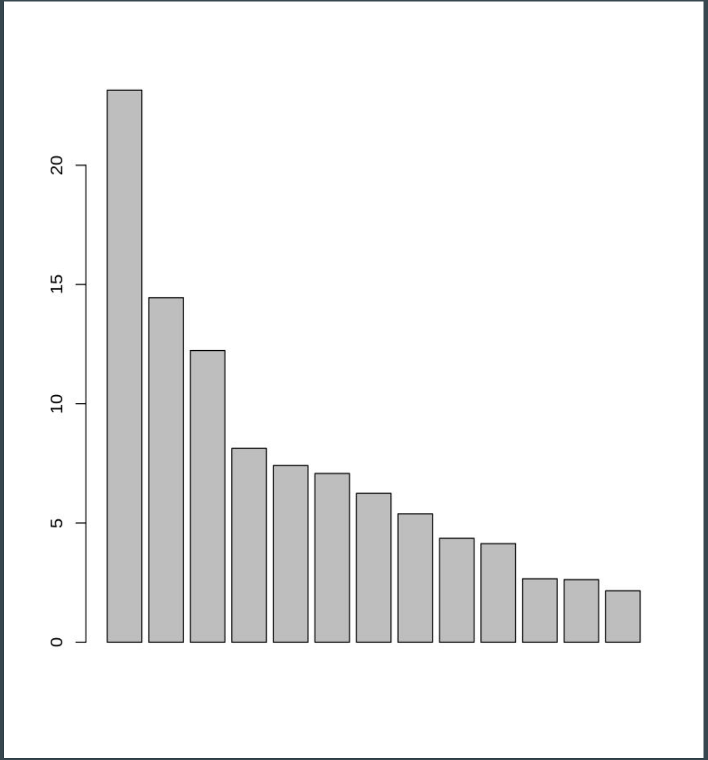
Interest correlation



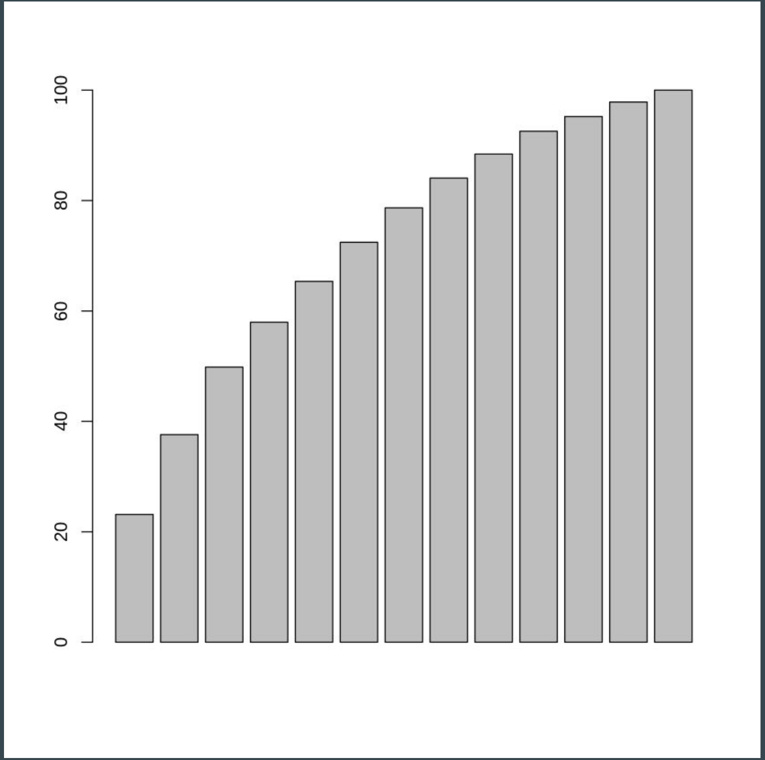
Numerical

PCA

Intertia by dimensions



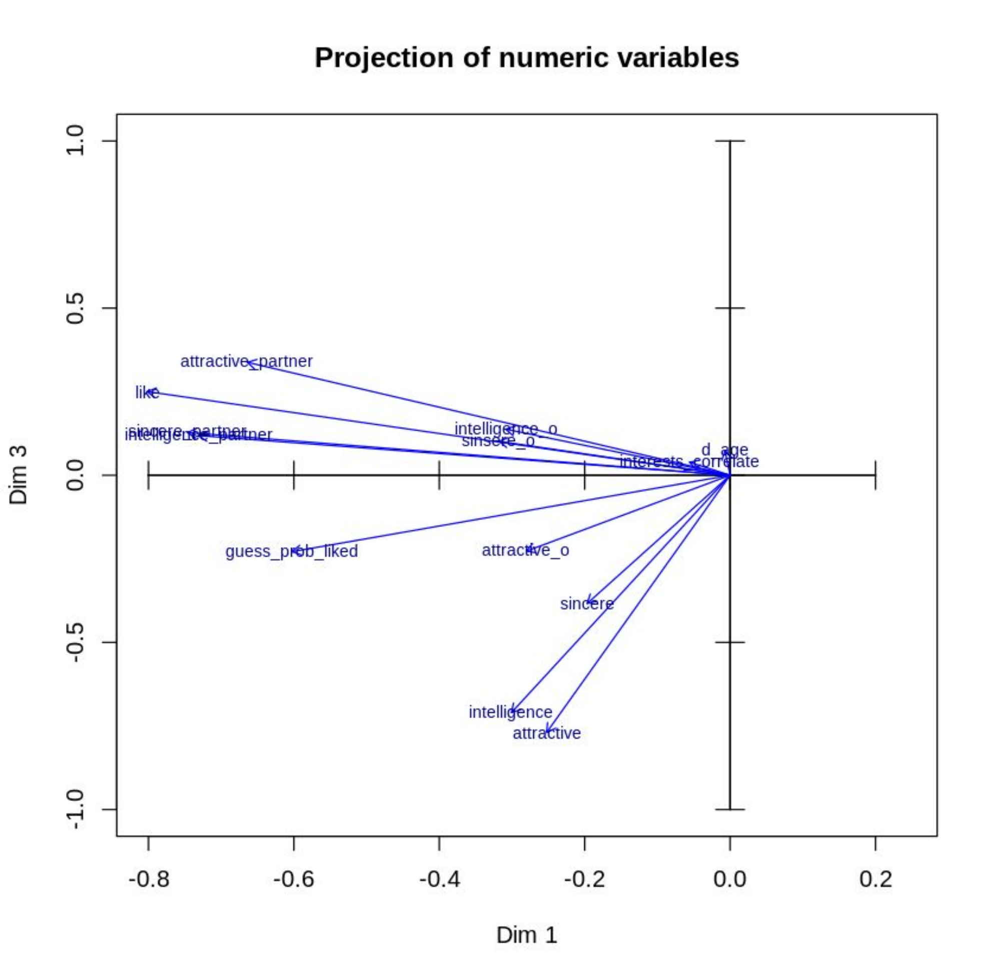
Acummulated inertia



PCA

Subspace
(1,3)

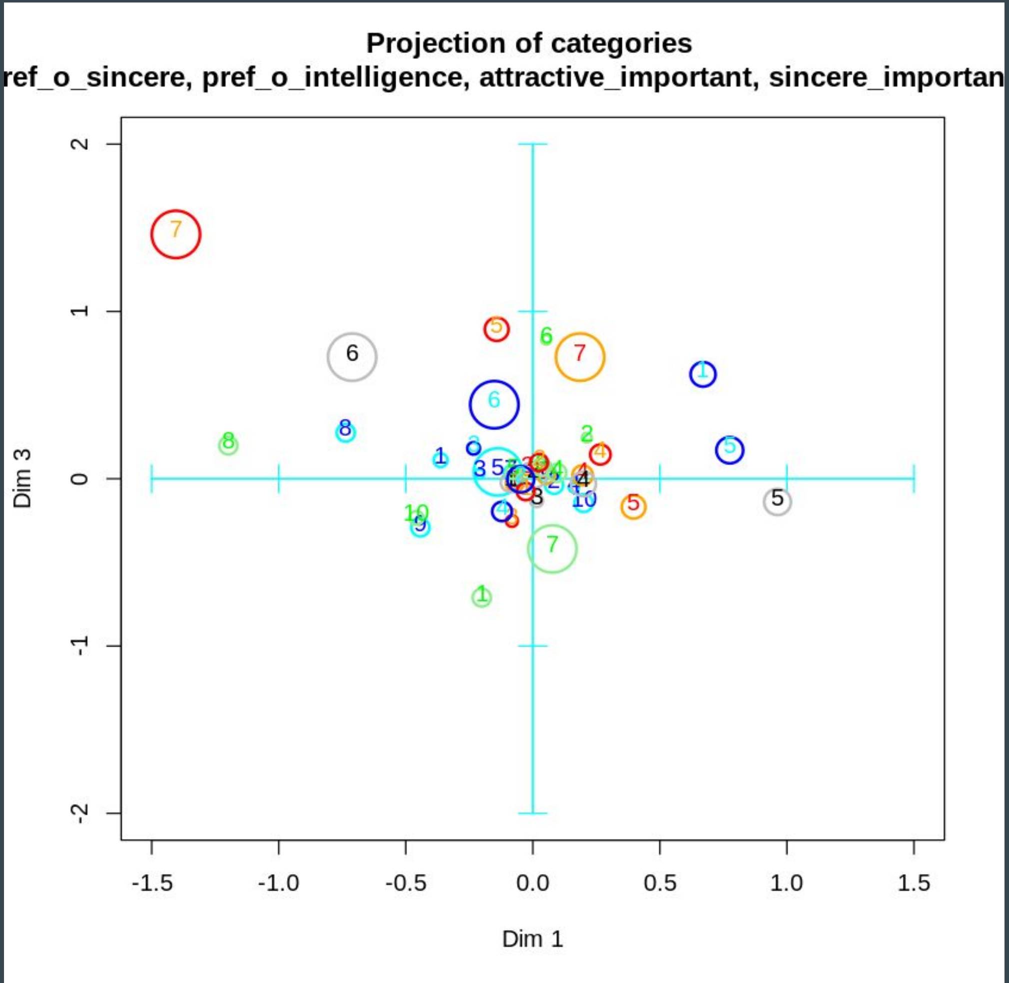
Projection of numerical variables



PCA

Subspace
(1,3)

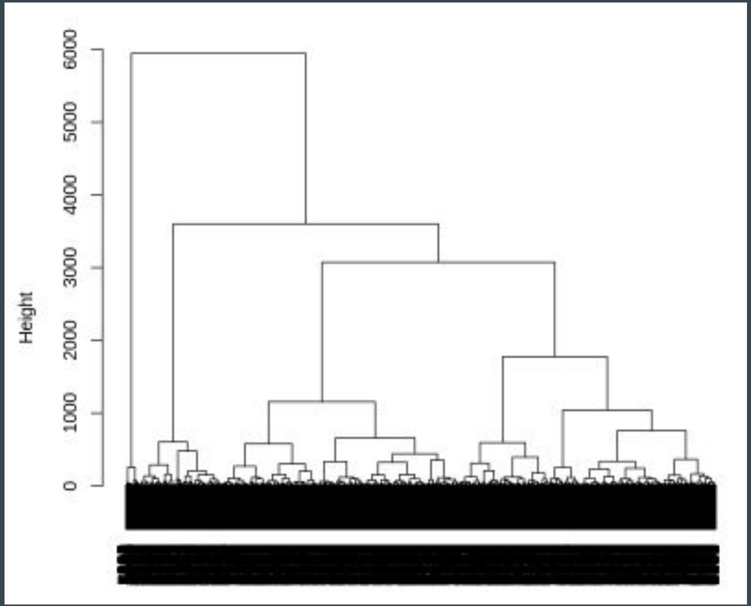
Projection of numerical and categorical variables



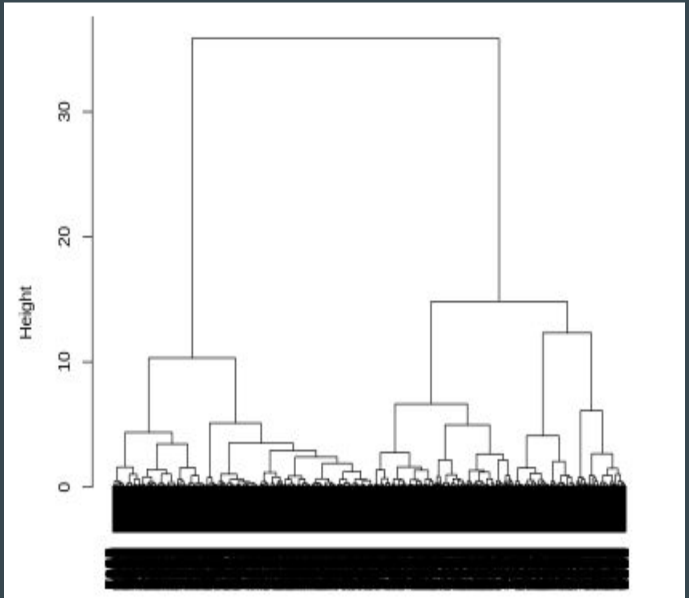
Clustering (only numerical variables)

Dendogram with distance

Dendogram with distMatrix

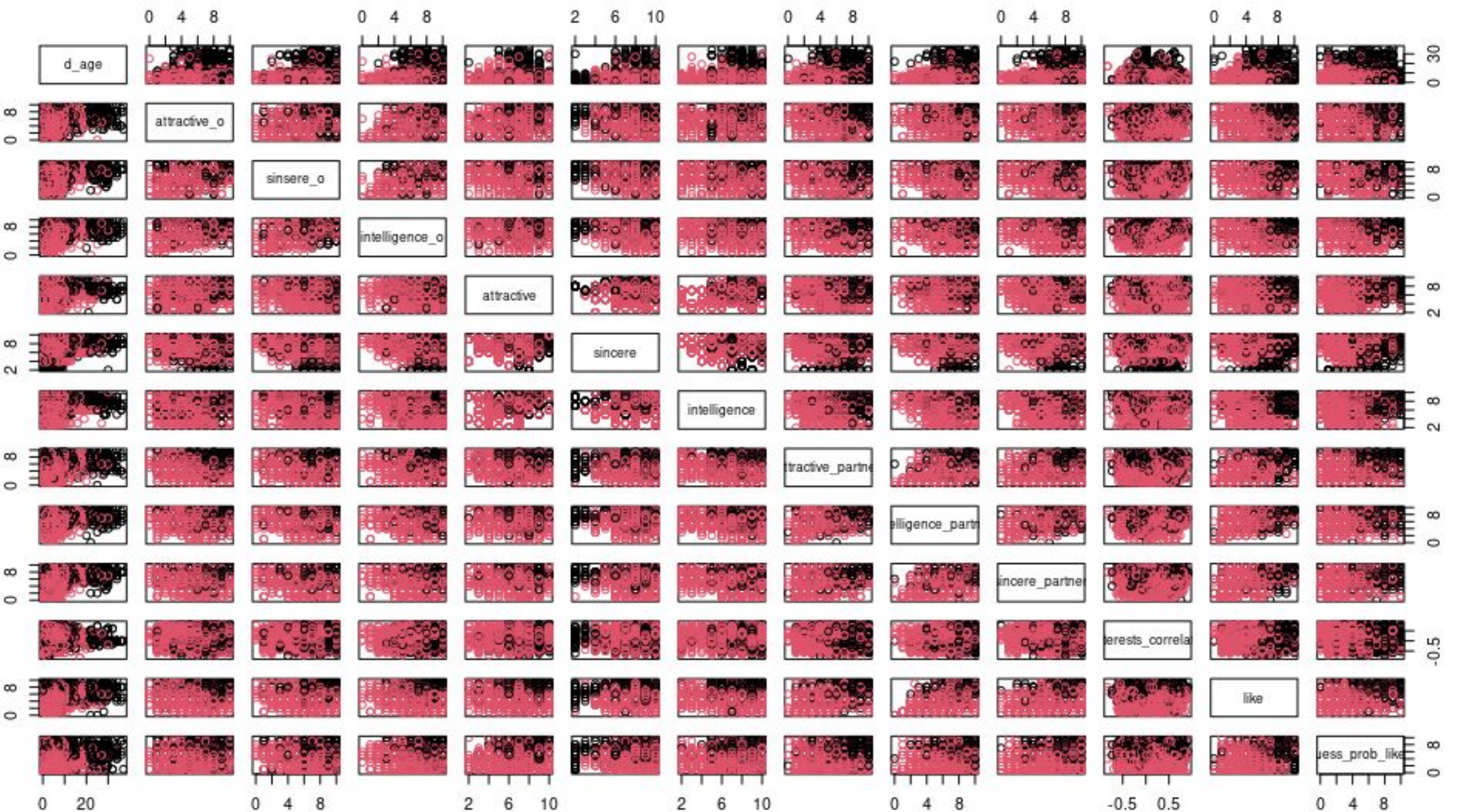


196 individuals of 8378 in c2

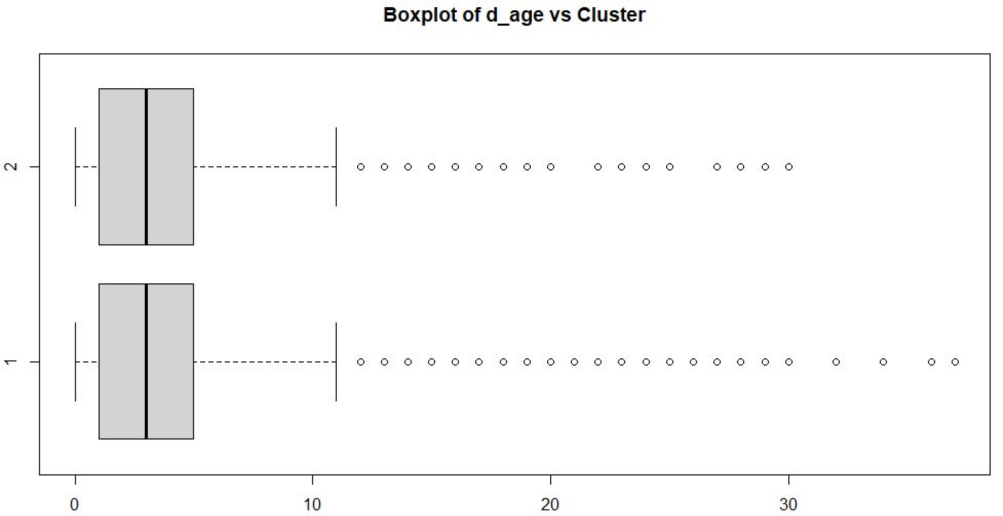


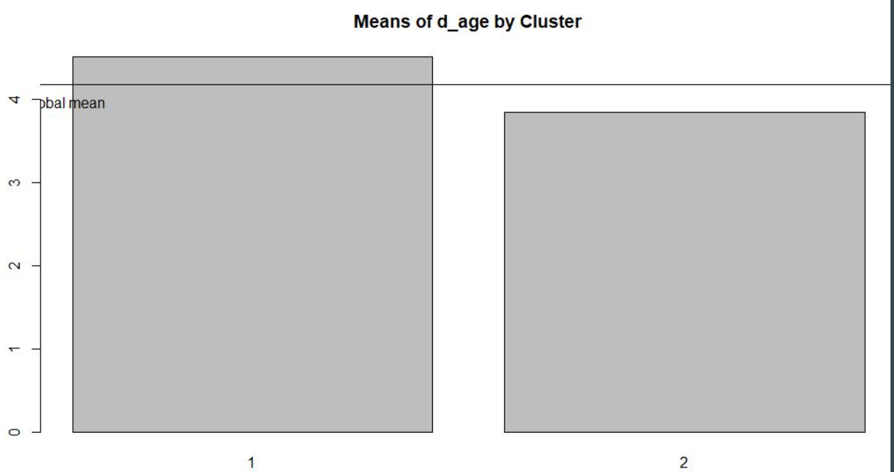
4112 individuals of 8378 in c2

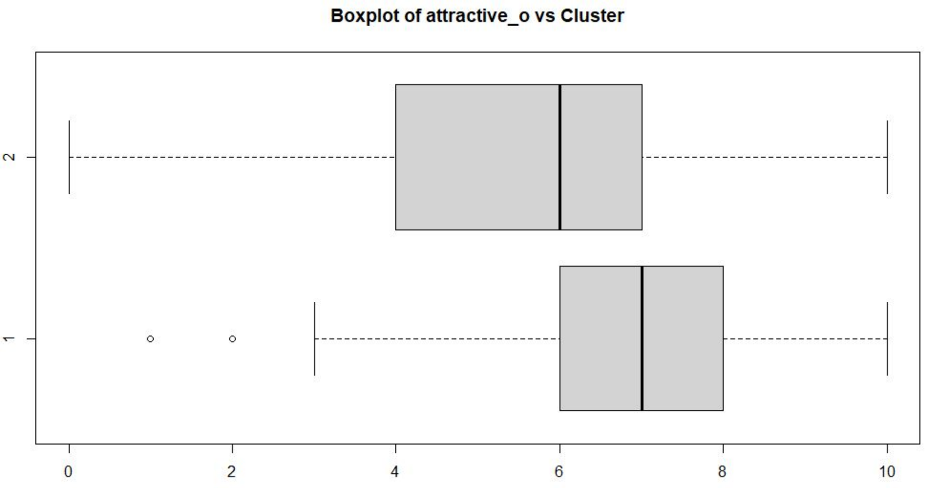
Clustering


d_age
attractive_o
sinsere_o
intelligence_o
attractive
sincere
intelligence
ttractive_partne
elligence_partn
incere_partner
erests_correla
like
uess_prob_like
0 4 8
0 4 8
2 6 10
0 4 8
0 4 8
0 4 8
30
0
8
0
8
2
8
2
8
0
-0.5
8
0
0 20
0 4 8
2 6 10
2 6 10
0 4 8
-0.5 0.5
0 4 8

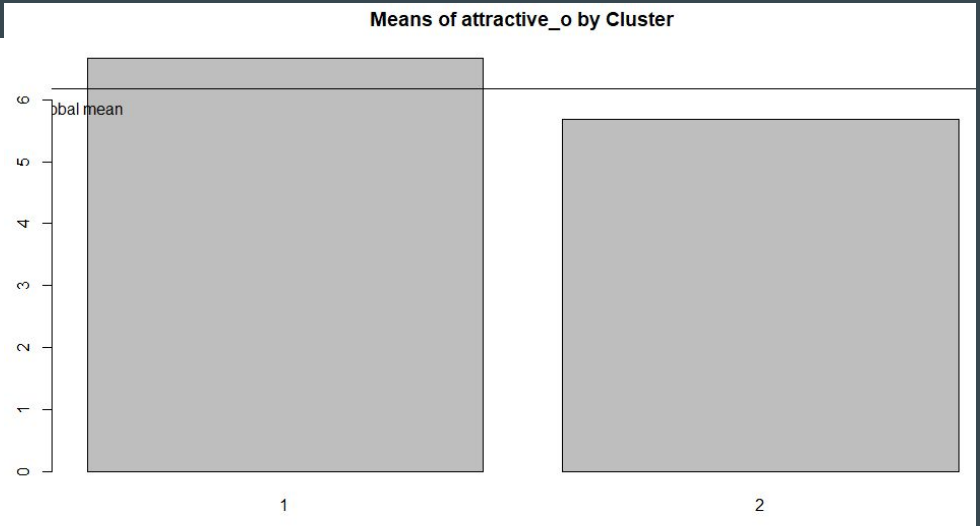
Profiling


Boxplot of d_age vs Cluster
2
1
0
10
20
30

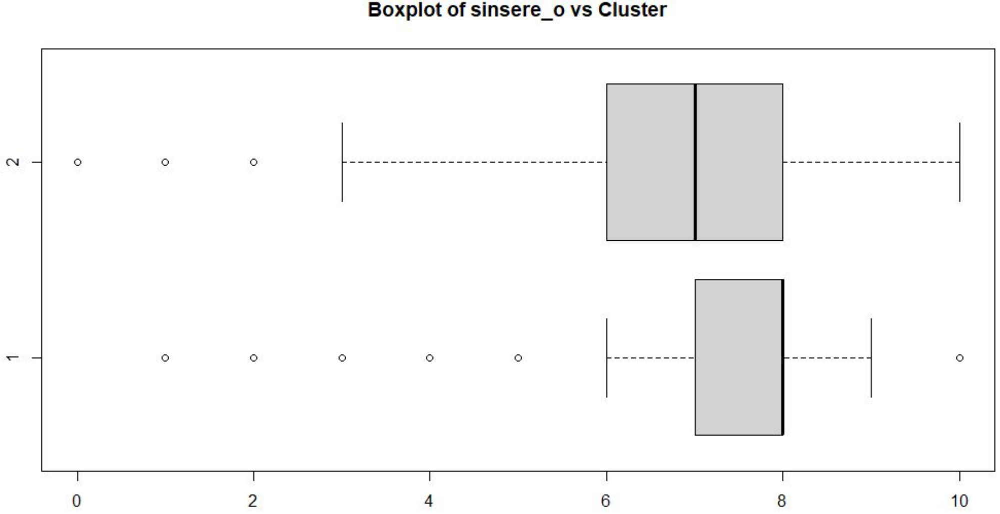

Means of d_age by Cluster
obal mean
0
1
2
3
4
1
2



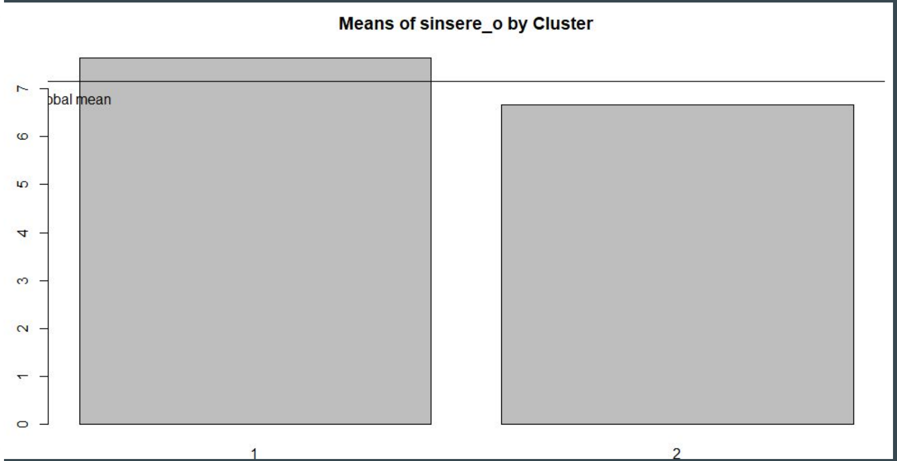
Boxplot of attractive_o vs Cluster



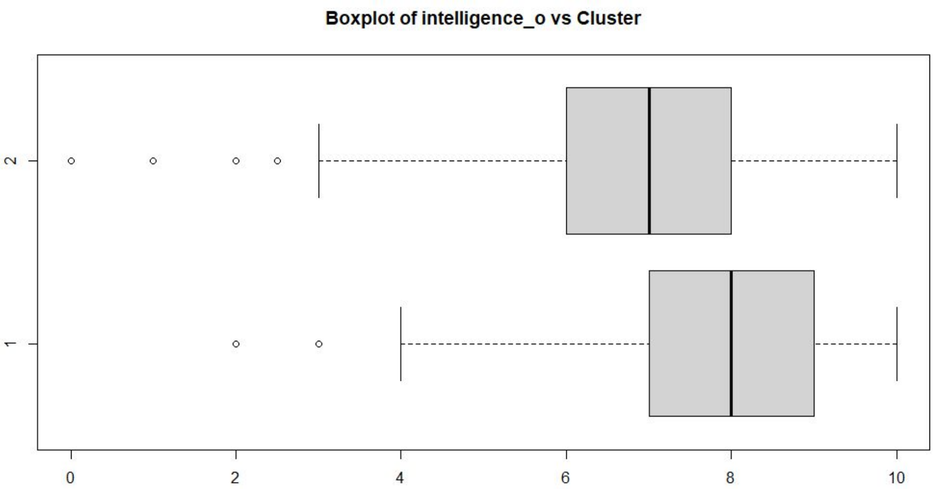
Means of attractive_o by Cluster

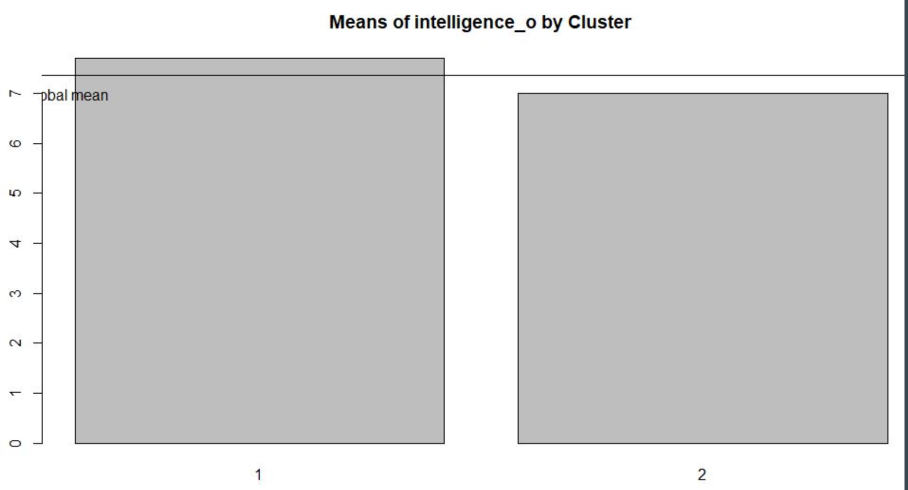


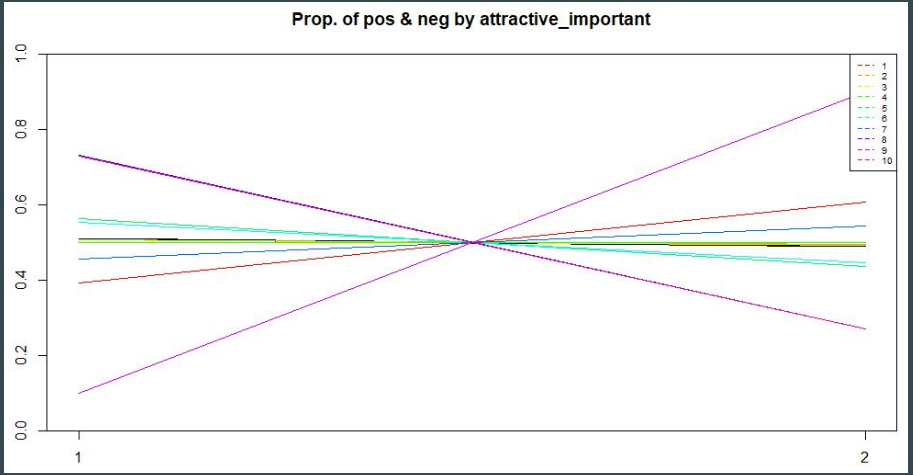
Boxplot of sinsere_o vs Cluster

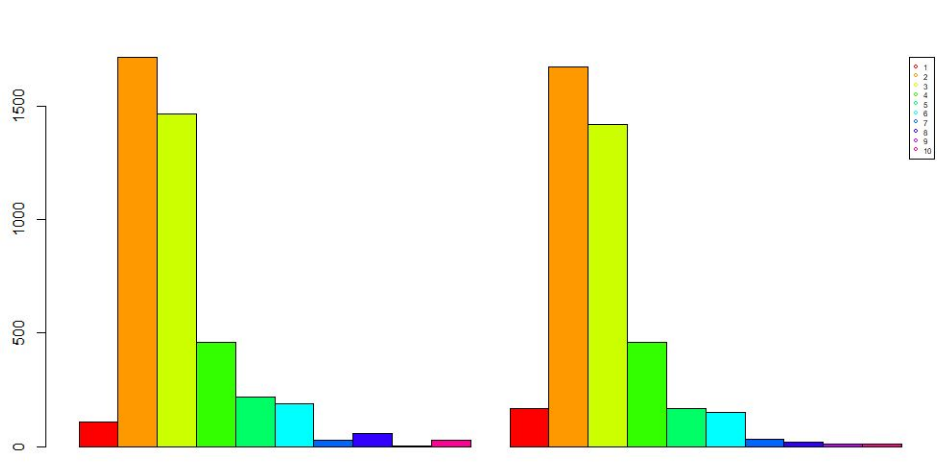


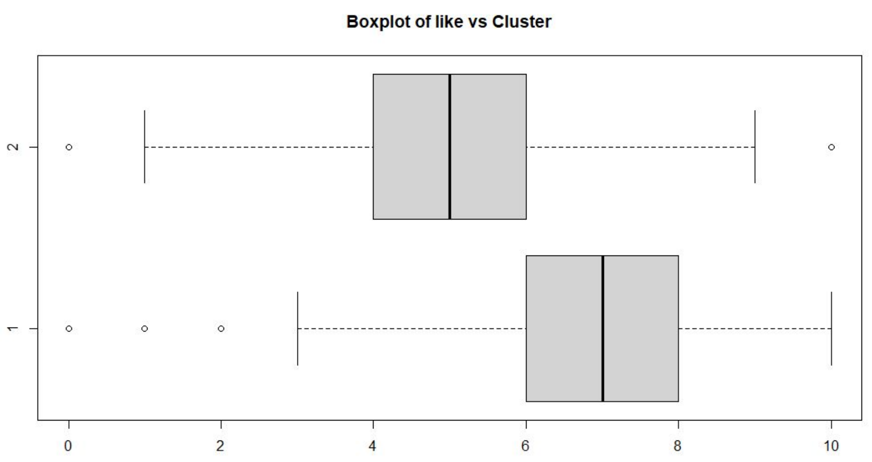
Means of sinsere_o by Cluster

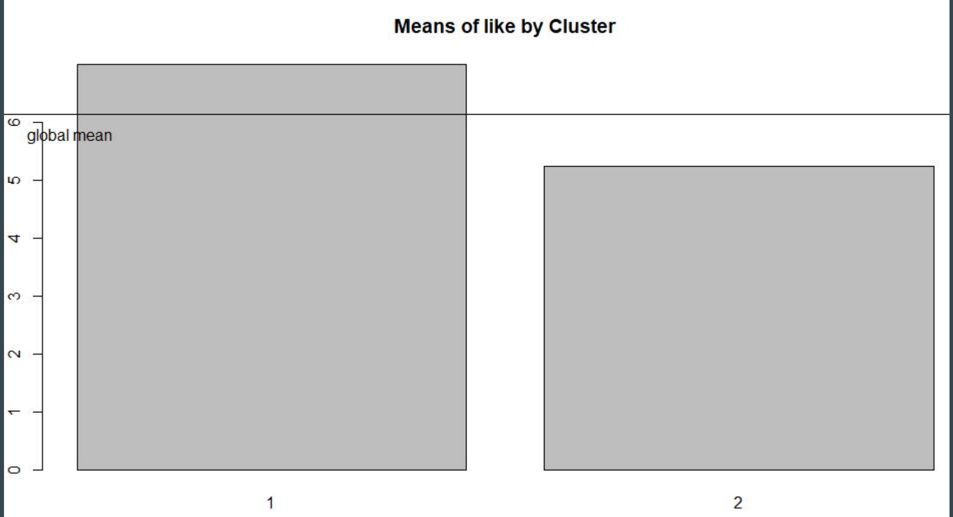

Boxplot of intelligence_o vs Cluster
2
1
0
2
4
6
8
10


Means of intelligence_o by Cluster
obal mean
0
1
2
3
4
5
6
7
1
2


Prop. of pos & neg by attractive_important
1.0
0.8
0.6
0.4
0.2
0.0
1
2
1
2
3
4
5
6
7
8
9
10


1500
1000
500
0
1
2
3
4
5
6
7
8
9
10

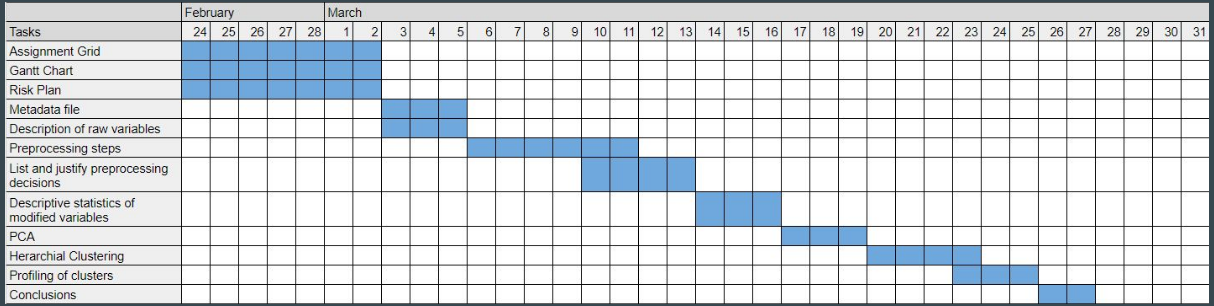

Boxplot of like vs Cluster
2
1
0
2
4
6
8
10


Means of like by Cluster
global mean
0
1
2
3
4
5
6
1
2

Conclusions

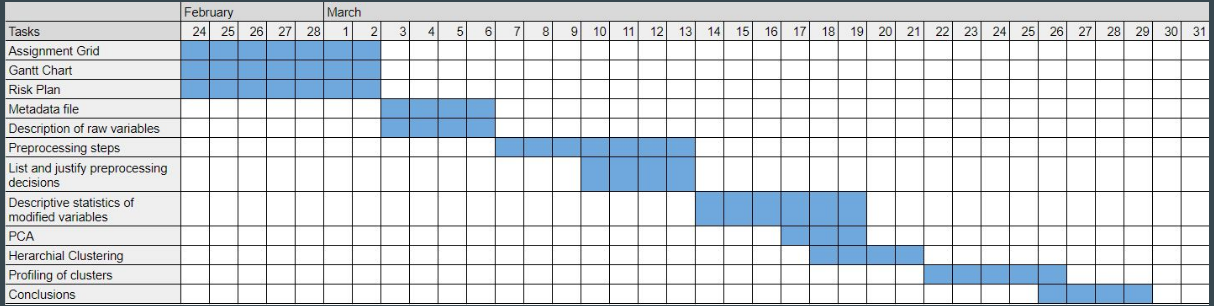
- Separability in how much their partners like them
- Intelligence > Sincere > Attractive

Original Scheduling



	February					March																													
Tasks	24	25	26	27	28	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Assignment Grid	■	■	■	■	■	■	■																													
Gantt Chart	■	■	■	■	■	■	■																													
Risk Plan	■	■	■	■	■	■	■																													
Metadata file								■	■	■																										
Description of raw variables								■	■	■																										
Preprocessing steps											■	■	■	■	■	■																				
List and justify preprocessing decisions															■	■	■	■																		
Descriptive statistics of modified variables																			■	■	■															
PCA																						■	■	■												
Herarchial Clustering																									■	■	■	■								
Profiling of clusters																												■	■	■						
Conclusions																															■	■				

Final Scheduling



	February					March																													
Tasks	24	25	26	27	28	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Assignment Grid	■	■	■	■	■	■	■																													
Gantt Chart	■	■	■	■	■	■	■																													
Risk Plan	■	■	■	■	■	■	■																													
Metadata file								■	■	■	■																									
Description of raw variables								■	■	■	■																									
Preprocessing steps												■	■	■	■	■	■	■																		
List and justify preprocessing decisions															■	■	■	■																		
Descriptive statistics of modified variables																			■	■	■	■	■	■												
PCA																						■	■	■												
Herarchial Clustering																							■	■	■	■										
Profiling of clusters																											■	■	■	■	■					
Conclusions																															■	■	■	■		

Questions ?